Statistiky využití článků v online repozitářích

Jan MACH
Vysoká škola ekonomická v Praze
Univerzita Karlova. Ústav informačních studií a knihovnictví
machj@vse.cz

INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích
Praha, 26. - 27. 5. 2015

Abstrakt:

V posledních letech vzniklo několik projektů, které se snaží využít metrik stávajících či najít metriky nové pro měření impaktu vědeckých informací publikovaných v režimu Open Access, tedy volně dostupných na Internetu. Vzhledem k charakteru Open Access publikování vědeckých informací se plný text práce nenachází v jednom časopise, není dostupný jen na jednom webu, ale stejný dokument může být dostupný na mnoha webových sídlech najednou a je na uživateli, jak a kde bude informace získávat. Příspěvek shrnuje nejdůležitější projekty zabývající se měřením impaktu publikování na Internetu a prezentuje projekty, které mají za cíl sběr, agregaci jednotlivých dílčích statistických dat a zpracování souhrnných indikátorů. V případě rozsáhlých Open Access repozitářů s významnými výsledky vědy a výzkumu nebo v případě publikace novinových článků na aktuální a nová témata se jako zajímavou alternativou klasických citačních metrik jeví alternativní metriky. Podmínkou využití alternativních metrik je atraktivnost dokumentu a jeho časté sdílení v sociálních sítích komunitou čtenářů. Vhodnost alternativních metrik autor ověřoval na příkladu repozitáře Pittsburské univerzity a statistik PlumX. V příspěvku je m.j. představen projekt IRUS-UK, který zajišťuje sběr základních dat ze spolupracujících repozitářů, která zpracovává do statistik odpovídajících standardu Counter. Poskytuje tak srovnatelné, autorizované a standardizované údaje o využití napříč repozitáři ve Velké Británii. Autor navrhuje obdobu projektu pro Českou republiku.

1 Úvod do metrik pro repozitáře

Vzhledem k prosazování Open Access publikování v otevřených repozitářích (např. ArXiv, NUŠL, v institucionálních repozitářích aj.) a Open Access časopisech (např. časopisy PLOS) vyvstává potřeba vhodných metrik pro ohodnocení významu vědeckých článků dostupných online v prostředí Internetu. Historicky byly vědecké práce spojeny s konkrétním časopisem, nyní jsou však často dostupné samostatně online a mohou být přečteny a využívány nezávisle na dostupnosti časopisu a jeho reputaci. Vědci již nečtou celé časopisy jako celek, ale elektronicky vyhledávají odpovídající články napříč více zdroji, korelace mezi významem časopisu a citovaností článku klesá (1).

V této práci autor hledá odpověď na otázku, jaké metriky jsou vhodné pro hodnocení článků dostupných v Open Access repozitářích. Ideální metrika by měla indikovat významnější práce v repozitáři již krátce po publikování a pomoci např. při třídění záznamů podle relevance.

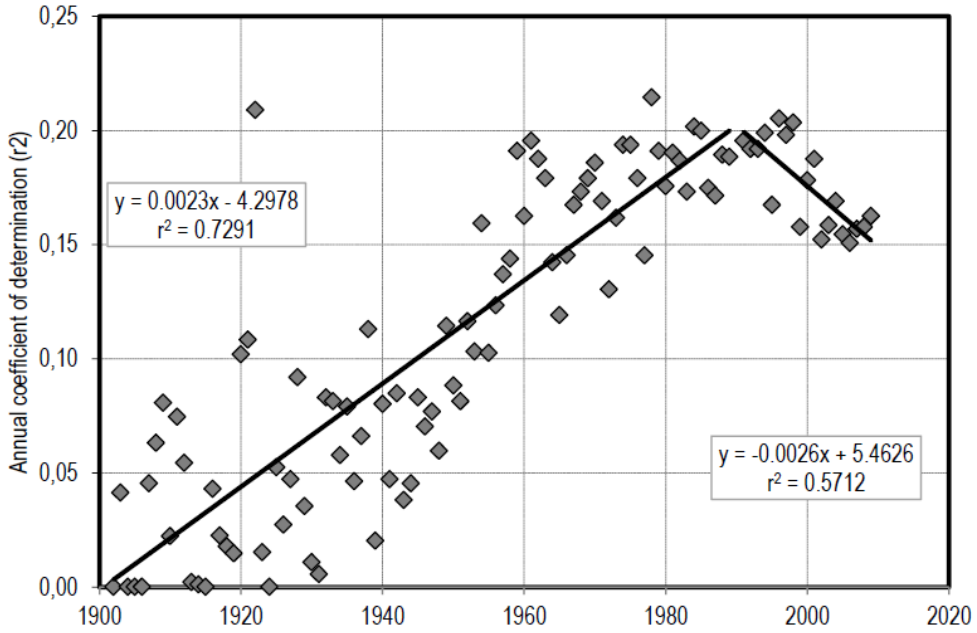
1.1 Metriky na bázi počtu citací

Nejznámějším indikátorem citovanosti časopisu založeným na počtu citací je Journal Impact Factor (zkráceně JIF), který poprvé zmínil Eugene Garfield v roce 1955. Vzhledem k vývoji citovanosti časopisů se JIF počítá pro každý konkrétní rok, je pak možné sledovat trendy ve vývoji JIF daného

periodika v čase. Nedostatky klasického Journal Impact Factoru řeší metriky z impakt faktoru odvozené, určené pro hodnocení časopisů, vědců, vědeckých týmů či institucí.

Hodnocením využití a dopadu konkrétních článků a autorů se zabývá citační analýza, konkrétněji metriky citační ohlas, Hirschův index a metriky z h-indexu odvozené.

Korelace mezi množstvím citací článku a impakt faktorem časopisu (viz Obrázek 1), ve kterém byl publikován, roste v průběhu 20. století a klesá po roce 1990 s nástupem digitálního věku a open access repozitářů, jak dokazuje studie (1).



Obrázek 1 Koeficient determinace r2 mezi IF časopisů z fyziky a počtem citací za 2 roky článků publikovaných v nich v letech 1902 až 2002 (1)

Studie dále prokazuje, že „podíl 10 % nejcitovanějších prací publikovaných v 10 % nejcitovanějších časopisů klesá od roku 1990, z 5,25 % na 4,50 %. V souladu s tím, podíl 10 % nejcitovanějších prací nezveřejněných v 10 % časopisů s nejvyšším impact faktorem od roku 1990 roste, z 52 % na cca 56 %. Tento vývoj je ještě zřetelnější, když je stejné srovnání zpracováno pro horních 5 % prací a horním 5 % časopisů.“ (1, str. 10). Vzhledem k závěrům, že citace se začínají více rozprostírat mezi jednotlivé časopisy, studie predikuje, že digitální věk a metody šíření a zpřístupňování vědeckých zdrojů mohou potlačit důležitost impakt faktoru jakožto významného kritéria důležitosti vědeckých publikací.

Varování před nevhodným užitím impakt faktoru a jeho problémy jsou uvedeny v mnoha pracích, důkladně se jim věnuje např. článek *Impakt faktory: využívání a zneužívání* autorů ze společnosti Elsevier (2).

1.2 WWW a Open Access

S nástupem World Wide Web se mění způsob vědecké komunikace, od klasického publikování v tištěných periodikách k šíření vědeckých informací v online prostředí Internetu, od publikování v časopisech kupovaných čtenáři k volnému přístupu placenému někým jiným než uživatelem – např. autorem, institucí apod.

Vyvstala tak potřeba nových metrik, nad rámec metrik na bázi impakt faktoru, které vezmou v potaz nové faktory úspěšnosti vědecké práce než jen pouhý počet citací v impaktovaných časopisech. Může se jednat o indikátory na bázi odkazů, kde odkaz na dokument v online repozitáři odpovídá citaci

v impaktovaném časopise, o indikátory známé z webometrie používané pro hodnocení WWW stránek nebo o tzv. sociální metriky, které měří odezvu dokumentu v sociálních sítích.

Dokumenty na Internetu získávají širší počet čtenářů než ty publikované v tištěných periodikách. Díky sledování webových aktivit (nepublikovaných materiálů, draftů, prezentací, studijních materiálů, komentářů, záložkování apod.) vědeckých skupin, profesorů, postgraduálních studentů aj. mohou vzniknout nové indikátory, postihující buď důležitost vědecké práce v odlišném ohledu než klasická citační analýza na bázi impakt faktoru, nebo kvality vědecké práce nepostihnutelné citační analýzou (práce kontroverzní nebo podněcující diskusi, horká témata, ekonomické, sociální, kulturní nebo oborové vztahy mezi autory apod.).

Z velkého množství webometrických indikátorů, které nám mohou pomoci při hodnocení významu vědeckých publikací, můžeme jmenovat např. sledování počtu odkazů na daný dokument (měří spíše atraktivitu dokumentu než jeho význam), viditelnost odkazů a jejich význam, měření velikosti webových sídel a počtu akademických prací v repozitářích nebo počty stažení.

1.2.1 Počet stažení

Počet stažení využívají v metrikách např. projekty COUNTER, PIRUS, LogEC aj. V praxi se odlišuje počet zobrazení vstupní HTML stránky s metadaty (např. jména autorů, název článku, abstrakt, klíčová slova), počet zobrazení plného textu (nejčastěji PDF), příp. i stažení sémantické reprezentace dokumentu (např. XML, Dublin Core, RDF). Takto odlišené indikátory můžeme najít např. v rámci Open Access časopisu *PLOS ONE* (3) viz Obrázek 2.

Article Usage

Total Article Views
943
Mar 27, 2012 (publication date) through undefined NaN, NaN*

	HTML Page Views	PDF Downloads	XML Downloads	Totals
PLoS	745	115	29	**889**
PMC	28	26	n.a.	**54**
Totals	**773**	**141**	**29**	**943**

Obrázek 2 Metriky PLOS ONE založené na počtu zobrazení článku (3)

Rizikem při použití indikátoru tohoto typu je nebezpečí generování a následného započítání falešných návštěv, a to ať již těch nechtěných – prostřednictvím vyhledávacích robotů, nebo záměrně automaticky generovaných samotnými autory pro umělé navýšení návštěvnosti.

Pro správnou interpretaci metrik je proto nutné ze statistik odstranit přístupy robotů a opakované klikání jednoho uživatele (tzv. „double clicks“). Často se používá negativního seznamu IP adres a jmen vyhledávacích robotů, příp. statistický přístup.

Indikátor počet stažení je jednoduché získat ze statistik webového serveru, pouze je nutné při návrhu webu dodržet zásadu jednoznačnosti a neměnnosti identifikace cílového dokumentu (URL adresy) a vhodně eliminovat vliv falešných stažení.

1.2.2 Návštěvnost, návštěvníci

Návštěvníkem je myšlena návštěva konkrétního uživatele po určitý časový úsek, tj. jeden návštěvník během své práce většinou stahuje ze stejného webu více dokumentů, příp. i stejný dokument opakovaně.

Nejednotnost měření počtu návštěvníků v jednotlivých repozitářích může být dána rozdílnou definicí časového úseku, což některé metodiky řeší jednotným, centralizovaným zpracováním institucionálních logů (např. projekt Open Access Statistics viz níže) nebo využitím centrálního logu

(např. u benchmarkingu knihoven BIX obsahuje cílová WWW stránka průhledný obrázek stahovaný z centrálního serveru, tj. logy jsou generovány a následně zpracovány jednou univerzitou). Pro knihovní prostředí je vhodné vycházet z definice virtuální návštěvy definované v *ISO 2789:2006 Information and documentation*, aby bylo možné porovnávat výsledky mezinárodně.

1.2.3 Sociální sítě, záložkování, citační manažery

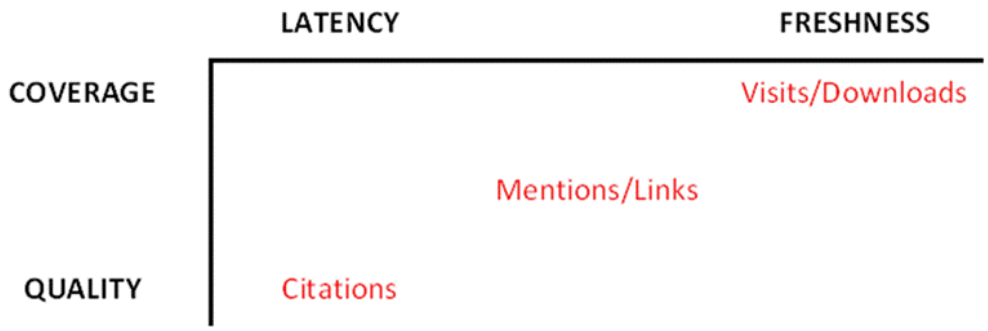
Uživatelé mohou v internetových aplikacích pro tzv. sociální záložkování (angl. social bookmarking) vytvářet záznamy o webových stránkách podobně, jako se vytvářejí záložky oblíbených stránek v prohlížečích WWW. Primárním účelem takovýchto aplikací, jejichž využití je již na ústupu, bylo nejen umožnit online přístup k záložkám, ale také snazší vyhledávání WWW stránek za využití citační analýzy (počet záznamů jednotlivých URL adres, počet a frekvence jednotlivých klíčových slov přiřazených uživateli). Příkladem služeb pro správu odkazů jsou reddit.com (http://reddit.com), del.ico.us (http://www.delicious.com) nebo služba zkracování odkazů Bitly (https://bitly.com).

Obdobný přínos jako záložkovací služby mohou pro nás mít sociální sítě na Internetu a jejich systém odkazování a hodnocení – blogy (odkazy, komentáře, hodnocení), odkazování ve Wikipedii a aktivity v sociálních sítích Facebook (odkazy, sdílení, komentáře, Like) či Twitter (odkazy, sledování, retweets).

Pro naši potřebu hodnocení Open Access článků mají pro nás o něco větší význam online citační manažery, které se více zaměřují na vědeckou komunikaci, využívány jsou spíše akademickou obcí. Kromě správy webových zdrojů je možné vytvářet i záznamy pro tištěné publikace a tyto záznamy dále kategorizovat. Příkladem citačních manažerů jsou RefWorks (http://www.refworks.com), ProQuest Flow (http://flow.proquest.com), EndNote (http://endnote.com), Mendeley (http://www.mendeley.com) či Zotero (http://www.zotero.org).

1.3 Alternativní metriky

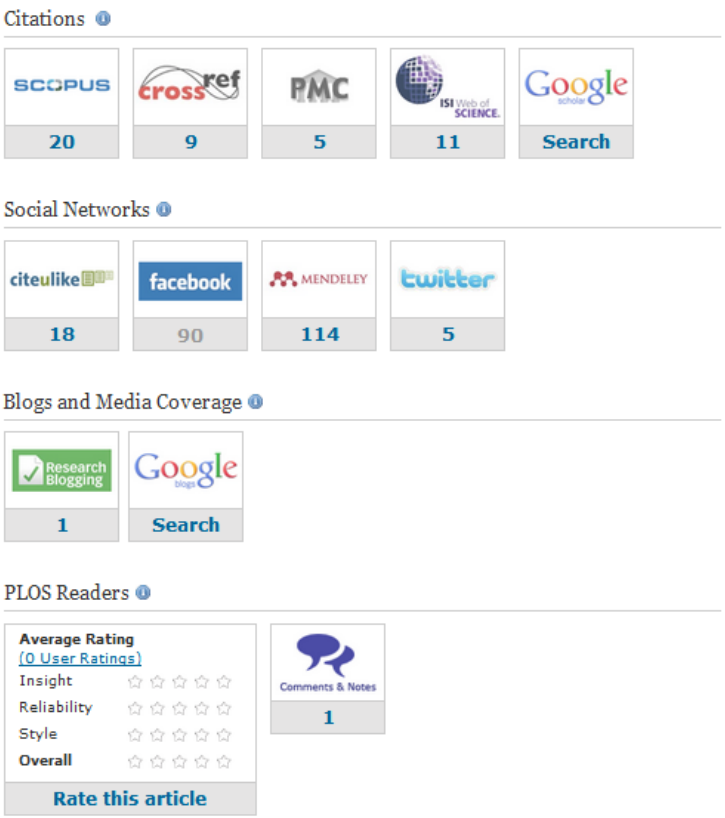
Nově navrhované alternativní metriky využívají (na rozdíl od tradičních metrik založených na analýze citací v tištěných vědeckých publikacích) měření událostí vyvolaných konkrétním článkem v sociálních sítích. Typicky se jedná o metriky založené na citacích/počtu odkazů na webových stránkách, počtu stažení, míře citace a aktivity na blozích a v dalších sociálních médiích. Sociální sítě reflektují nové publikace rychleji než klasické citování v tištěných publikacích, zahrnují různé typy vědeckých výstupů i impakt na neakademickou sféru, a to v daleko větší míře nuancí než běžné metriky citační analýzy. Rozdíl mezi klasickou citační analýzou tištěných periodik a novými metrikami na bázi webometrik a sociálních sítí výstižně zobrazuje následující schéma projektu OpenAIRE (4) viz Obrázek 3. Zatímco metriky založené na měření počtu citací jsou vhodné pro kvalitní články, ale důvěryhodné výsledky poskytují až po delší době po publikování, metriky založené na měření počtu návštěv a stahování jsou určeny pro široké pokrytí různých online zdrojů, s možností vyhodnocení již krátce po publikování dokumentu na webu.



Obrázek 3 Vliv jednotlivých metrik na pokrytí, kvalitu a reakční čas (4)

Mezi jednu z nejvýznamnějších iniciativ zaměřených na nové metriky pro měření impaktu Open Access publikací patří projekt Article Level Metrics – altmetrics, jehož autoři stanovili potřebu nových metrik ve svém programovém prohlášení *altmetrics manifesto* (5).

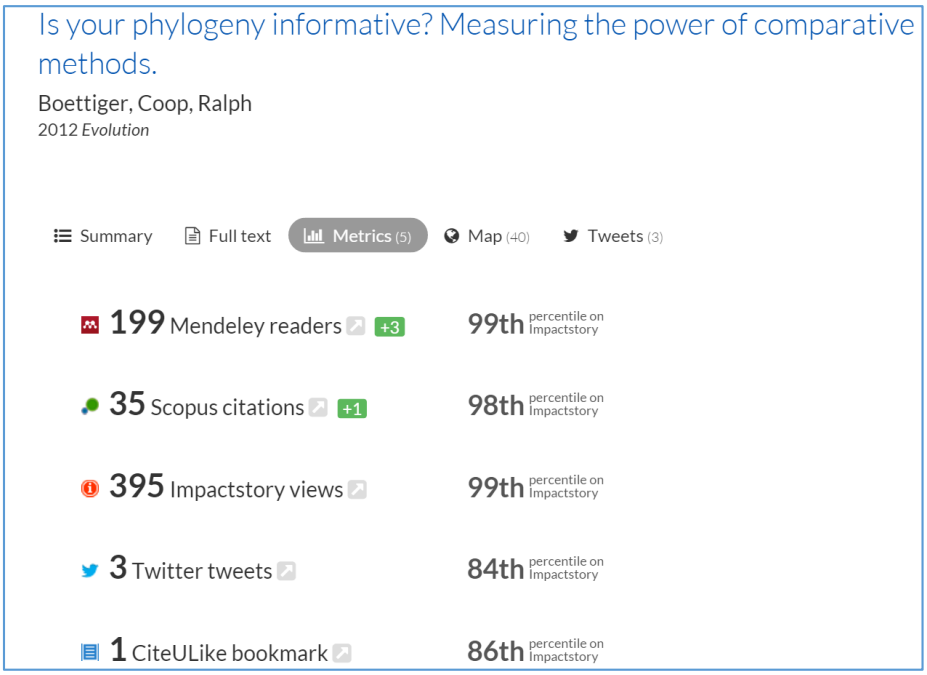
Příkladem aplikace alternativních metrik je časopis *PLOS ONE* (3), recenzovaný vědecký časopis zaměřený na primární výzkum, vydávaný Public Library of Science (zkratka PLOS, původně PLoS). Časopis *PLoS ONE* byl spuštěn v prosinci 2006 s funkcionalitou komentování a tvorby poznámek. Později přibyly funkce hodnocení článků, zpětných odkazů (trackbacks). Vydavatel PLoS začal v souladu s projektem Article Level Metrics zveřejňovat online uživatelská data pro publikované články, v současnosti *PLOS ONE* poskytuje řadu indikátorů včetně citačních metrik, statistik využití, pokrytí v blozích, komunitního záložkování, komunitního a expertního hodnocení (viz Obrázek 4). Využití jednotlivých indikátorů pro zhodnocení kvality a impaktu článku na základě jednotlivých metrik je již ponecháno na čtenáři.



Obrázek 4 Metriky PLOS ONE založené na citační analýze, sociálních sítích a hodnocení čtenářů (3)

Webový nástroj Impactstory (http://www.impactstory.org) (6), původním jménem Total-Impact, umožňuje uživatelskou tvorbu kolekcí dokumentů na základě různých zdrojů – seznamů DOI, účtů SlideShare apod. Pro nalezené dokumenty (články, prezentace, postery, data z výzkumů apod.) Impactstory spočítá a zobrazí metriky podle altmetrics.

Jednotlivé dokumenty v Impactstory profilu autora mohou být, na základě automatického porovnání s ostatními dokumenty na Impactstory, ohodnoceny popisky typu (highly) cited, downloaded, discussed, recommended anebo viewed. Pro jednotlivé metriky (např. čtenáři Mendeley, citace Scopus, zobrazení v Impactstory, záložky v CiteULike aj.) je zobrazen trend vývoje za poslední období a percentile na Impactstory pro daný rok publikace dokumentu (viz Obrázek 5).



Obrázek 5 Metriky článku na webu Impactstory (6)

Poslední ze zde uvedených nástrojů na měření dopadu vědy je nástroj PlumX od Plum Analytics (https://plu.mx) (7), který na základě sběru dat z jednotlivých zdrojů nabízí metriky pro měření vlivu publikací, vědců, vědeckých skupin a institucí. Nástroj PlumX poskytuje widgety a API pro přidání jednotlivých metrik do institucionálních repozitářů, využití poskytovaných metrik je opět na samostatném uživateli.

2 Agregace a zpracování statistických dat

Vzhledem k charakteru Open Access publikování vědeckých informací se plný text práce nenachází v jednom časopise, není dostupný jen na jednom webu, ale stejný dokument může být dostupný na mnoha webových sídlech najednou a je na uživateli, jak a kde bude informace získávat. Aby bylo možné hodnotit i takovýto způsob publikování, je potřeba

a) jasně identifikovat dokument pomocí perzistentních identifikátorů, např. DOI (viz např. projekt COUNTER dále), urn:nbn, PURL, HANDLE aj. tak, aby bylo možné jednotlivé statistiky agregovat (problematika perzistentních identifikátorů přesahuje rámec této práce),
b) zajistit sběr, agregaci a zpracování jednotlivých dílčích statistických dat o využití jednotlivých dokumentů,
c) standardizovaný výpočet souhrnných indikátorů ze získaných dat.

Pokud získáme seznam záznamů konkrétní práce z uvedených aplikací (ať již identifikovaných podle URL nebo přesněji podle DOI aj. identifikátorů), můžeme využít nástrojů citační analýzy podobně jako při měření citačního indexu pro tištěné publikace.

2.1 Counter, Sushi

Nejznámějším standardem pro statistiky využití elektronických informačních zdrojů je standard *Counting Online Usage of Networked Electronic Resources* (zkráceně COUNTER) široce přijímaný jak producenty, tak kupujícími, knihovníky a dalšími zainteresovanými stranami. Standard definuje používané termíny, sadu metrik a reporty[1] evidující využití elektronických zdrojů – časopisů, databází, knih a multimédií v prostředí Internetu.

Mezinárodní sada pravidel a protokolů pro zaznamenávání a sdílení dat o využití elektronických informačních zdrojů je popsána v *Counter Code of Practice for eResources* (8). Verze 4 je platná od roku 2012, v roce 2014 byla rozšířena mj. o statistiky COUNTER měřící užití vědeckých zdrojů dostupné i na úroveň článku. *COUNTER Code of Practice for Articles* (9) definuje dvě zprávy, *Article Report 1* (počet všech úspěšných požadavků na článek, podle autora a měsíce) a *Article Report 2* (počet úspěšných požadavků na stažení plného textu článku podle autora, měsíce a DOI, sdružené podle zdrojů).

V poslední revizi *Counter Code of Practice* je pro nás významné, že byla mj. nově stanovena povinnost:

1) uvádět v reportech perzistentní identifikátor DOI pro časopisy a knihy za účelem lepší správy statistických dat a možnost propojení na online kolekce,
2) uvádět využití Open Access fulltextových článků v časopisech v samostatném reportu *Journal Report 1 Gold Open Access* (převzato z projektu SUSHI, viz dále, bohužel reporty poskytovány pouze na úrovni časopisů, nikoliv článků),
3) uvádět počet zobrazení fulltextového záznamu z výsledku, naopak povinnost statisticky vykazovat vágněji definované návštěvy (sessions) a celkové počty vyhledávání byla zrušena.

Automatizované stahování statistických dat je umožněno prostřednictvím protokolu SUSHI. Protokol SUSHI sám o sobě neřeší agregaci statistických dat pro stejný zdroj v různých databázích, ale díky nové povinnosti uvádět DOI ve statistikách COUNTER by toto mohl řešit např. systém správy elektronických informačních zdrojů.

2.2 PIRUS, PIRUS2

Program *Publisher and Institutional Repository Statistics* (zkráceně PIRUS) z roku 2008 sponzorovaný JISC měl za úkol vytvořit a upravit standardy, statistiky a procesy za účelem rozšíření standardu COUNTER na úroveň časopiseckých článků. Projekt demonstroval možnost tvorby, zaznamenávání a agregace takovýchto statistik využití, navržené XML schéma pro statistiky využití na úrovni jednotlivých článků bylo implementováno např. provozovateli repozitářů PLOS a SURF.

Navazující program PIRUS2 probíhal od listopadu 2009 do prosince 2010, jeho cílem bylo mj. ověřit škálovatelnost navržených metod, stanovit náklady na tvorbu reportů a centralizovanou správu a rozšířit akceptaci navržených metod mezi další provozovatele repozitářů, vydavatele a autory.

[1] Seznam reportů COUNTER ke stažení přes protokol SUSHI je k dispozici na http://www.niso.org/workrooms/sushi/reports/ .

Projekt PIRUS2 (10) doporučuje identifikovat zdrojové články pomocí metadatového balíčku OpenURL ContextObject pro časopisecké články[2]. V rámci projektu byly připraveny moduly pro repozitáře na bázi DSpace, EPrints a Fedora pro zpřístupnění dat o využití časopiseckých článků. Implementaci takto ověřené možnosti zpřístupňování a zpracování statistik na úrovni článků do protokolu COUNTER projekt doporučuje k dalšímu výzkumu.

Projekt PIRUS2 navrhl rozšíření projektu COUNTER o report *Article Report 1* – počet úspěšných požadavků na plný text článků podle DOI a měsíce, ve formátu Excel (ruční stažení) a XML (automatické zpracování, protokol SUSHI). Jako hlavní způsob konsolidace dat od jednotlivých poskytovatelů je navrženo využití perzistentního identifikátoru DOI.

Projekt PIRUS dále doporučil ustanovení centrální clearingové instituce (tzv. Clearing House) zajišťující sběr, konsolidaci dat a distribuci reportů o využití a navrhl ekonomický model provozu. Při události stažení plného textu je dle projektu přiřazen události kód (tzv. tracker code) a vygenerován OpenURL záznam v logu. Projekt PIRUS původně navrhoval 3 metody následného zpracování logů:

A. OpenURL záznam je zaslán na lokální server, kde je filtrován na základě pravidel COUNTER (eliminování robotů a tzv. dvojkliků), vygenerovány statistky využití COUNTER pro článek identifikovaný DOI v XML formátu a jejich následné zpřístupnění (vhodné pro velké repozitáře a producenty, preferováno)
B. OpenURL záznam je uložen na lokální OAI-PMH server, odkud je stažen protokolem OAI-PMH do Clearing House, kde je provedeno zpracování podle pravidel COUNTER a vygenerování statistik obdobně jako v případě lokálního serveru ve scénáři A, statistiky jsou následně zpřístupněny autorizovaným institucím.
C. OpenURL záznam je zaslán do Clearing House, kde je provedeno zpracování podle pravidel COUNTER a vygenerování statistik obdobně jako v případě lokálního serveru ve scénáři A, statistiky jsou následně zpřístupněny autorizovaným institucím.

Oproti návrhu projektu PIRUS scénáře B na využití protokolu OAI-PMH pro harvestování statistických dat (pro srovnání – OAI-PMH používají projekty Open Access Statistics a KE/SURFsure viz níže), projekt PIRUS2 klade důraz na zbývající dvě navržené metody zpracování dat a možnost vystavení dat pomocí OAI-PMH ponechává jako doplňující. (10)

V případě užití scénářů B a C při předávání a zpracování podkladových statistických dat je jednodušší agregace dat z různých repozitářů, využitím centrálního zpracování na základě pravidel COUNTER je zaručen konzistentní způsob zpracování nad všemi zapojenými repozitáři. Vzhledem k nižším nárokům na implementaci na straně lokálních repozitářů by implementace měla být jednodušší. (11)

2.3 Open Access Statistics

Problém chybějící standardizace v oblasti metrik využití vědeckých článků v prostředí Internetu se rozhodli řešit autoři, především z řad německých knihoven, projektu Open Access Statistics (http://www.dini.de/projekte/oa-statistik; zkráceně OAS). Projekt vychází ze standardů COUNTER, LogEc (agregace statistik přístupů služeb RePEc, na úrovni článků) a IFABC (metriky pro WWW servery, dokumenty, e-maily aj.).

Příkladem problémů řešených v projektu je přenos podkladových dat/logů a deduplikace uživatelů/dokumentů nejen v rámci jednoho repozitáře, ale také v síti zapojených Open Access repozitářů, kdy jeden uživatel může volně přecházet mezi různými repozitáři s identickými verzemi dokumentů. Centrální zpracování dat umožňuje také sledovat trend stahování různých dokumentů

[2] Bližší popis metadatových prvků viz http://ocoins.info/cobg.html.

jedním uživatelem z různých repozitářů, příp. součet požadavků na různé dokumenty příbuzného zaměření v odlišných repozitářích.

V rámci projektu OAS byla v první fázi (květen 2008 – prosinec 2010) vybudována síť repozitářů za účelem sběru a výměny informací o využití jednotlivých časopiseckých článků. V druhé fázi (duben 2011 – duben 2013) se projekt OAS, resp. OAS2 zaměřil na standardizaci a ověření indikátorů na bázi absolutní frekvence využití dokumentů, standardizaci procesů, uložení dat, rozhraní pro výměnu dat o využití, zajištění trvalé udržitelnosti a na integraci nových služeb, jako je např. řazení na bázi frekvence stahování, indikace impaktu dokumentu aj.

2.4 KE Usage Statistics Group, SURFsure

Iniciativa Knowledge Exchange (KE) má za cíl podporovat spolupráci mezi národními institucemi v Evropě odpovědnými za vývoj infrastruktury a služeb na podporu ICT ve vědě a výzkumu. V rámci aktivit zaměřených na spolupráci digitálních repozitářů, konkrétně na jejich statistiky využití, uspořádala řadu seminářů, kterých se účastnili odborní zástupci souvisejících projektů COUNTER, PIRUS, OAS a SURF Statistics on the Usage of Repositories (SURFsure, projekt na duben 2009 – březen 2010), jakož i např. zástupci služeb RePec a NeeO zaměřených na Open Access v ekonomii.

Vzhledem k absenci jednotného standardu sdílení statistik o využití mezi jednotlivými systémy a s tím spojených problémů, byl v rámci pracovní skupiny KE připraven k podzimu 2010 dokument *KE Usage Statistics Guidelines* (11), který vycházel z již probíhajících prací v souvisejících projektech. Tyto pokyny jsou nyní součástí projektu SURFsure.

Využití dokumentu je definováno jako zobrazení plného textu dokumentu nebo jemu přidružených metadat. Podobně jako je definováno v projektu PIRUS ve scénáři B (a využito např. v rámci OAS), pro stažení statistických dat clearingovým centrem je využito protokolu OAI-PMH nebo je možné použít protokol SUSHI viz scénář C projektu PIRUS.

V rámci *KE Usage Statistics Guidelines* je kromě způsobu přenosu také řešeno:

a) jednotný formát přenášených dat na bázi OpenURL Context Object, jehož vhodnost byla prokázána v předešlých projektech,
b) normalizace dat – filtrování „double clicks“ v clearingovém centru (opakované požadavky jedním uživatelem) a filtrování přístupů robotů v lokálních repozitářích a příp. pokročilejší heuristikou v clearingovém centru (seznam robotů je vytvořen kombinací seznamů z projektů COUNTER, AWStats, Universidade do Minho a PLOS),
c) pseudo-anonymizace dat – IP adres v souladu se *Směrnicí 95/46/ES o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů*.

Na základě výše uvedených pokynů byla zpracována mj. Pravidla OpenAIRE určená pro získávání statistik využití ze zapojených Open Access repozitářů v rámci pilotního projektu otevřeného přístupu v 7. rámcovém programu EK (4).

2.5 Projekt IRUS-UK

Za finanční podpory JISC byl vybudován projekt IRUS-UK (12), který vychází z doporučení stanovených v projektech PIRUS. V rámci projektu byla vybudována centrální clearingová instituce (tzv. Clearing House) zajišťující sběr, konsolidaci a distribuci reportů podle statistik Counter o počtu stažení záznamů z institucionálních repozitářů Velké Británie. Projekt byl prezentován Paulem Needhamem na konferenci ETD 2014 v Leicesteru, UK (13).

IRUS-UK zajišťuje sběr základních dat ze spolupracujících repozitářů, která zpracovává do statistik odpovídajících standardu Counter, resp. vychází z doporučení *COUNTER Code of Practice* (8) a *PIRUS Code of Practice for recording and reporting usage at the individual article level* (14). Poskytuje tak srovnatelné, autorizované a standardizované údaje o využití napříč repozitáři ve Velké Británii.

Konkrétní postup implementace pro IRUS-UK je popsán v dokumentu *IRUS-UK Code of Practice*, který definuje technické, organizační a ekonomické modely pro záznam, vykazování a sběr statistik využití všech typů záznamů zpřístupněných ve spolupracujících institucionálních repozitářích ve Velké Británii.

Sběr dat z institucionálních repozitářů probíhá podle specifikace *The Tracker protocol V3.1* (15). Podle této specifikace, pokud uživatel klikne na odkaz pro stažení souboru z repozitáře, je odeslán na vzdálený server k dalšímu zpracování OpenURL záznam ve formátu ContextObject (podle NISO standardu OpenURL 1.0). OpenURL řetězec obsahuje údaje ve formátu klíč=hodnota oddělené znakem & (viz *Tabulka 1*).

Tabulka 1 Prvky IRUS-UK Push protokolu (zdroj: vlastní zpracování podle (15)*)*

Prvek	OpenURL klíč	PZopis
OpenURL version	url_ver	Identifikace dat podle OpenURL 1.0
Usage event datestamp	url_tim	Časové razítko události (datum a čas)
Client IP address	req_id	IP adresa klienta požadujícího článek
UserAgent	req_dat	Řetězec UserAgent identifikující klientský program
Item OAI identifier	rtf.artnum	OAI identifikátor
FileURL	svc_dat	URL na plný text
HTTP Referer	rfr_dat	Pole HTTP Referer podle HTTP protokolu
Source repository	rfr_id	Identifikátor repozitáře, ve kterém došlo k události

Následně jsou záznamy v centrálním registru deduplikovány podle pravidel 5. sekce 4. revize *COUNTER Code of Practice* (8) o záznamy generované roboty (seznam robotů podle Counter doplněný o vlastní seznam IRUS-UK) a záznamy generované neobvyklým stahováním dat (dvojkliky a většinou více jak 100 stažení z repozitáře z jedné IP adresy denně).

Na základě zpracovaných dat je generována řada standardizovaných statistik, které jsou dostupné zapojeným institucím na webu projektu IRUS-UK.

3 Evaluace vhodnosti alternativních metrik pro Open Access repozitáře

Ukazuje se, že klasické citační metriky neposkytují dostatečné podklady pro evaluaci publikační činnosti v režimu Open Access, na úrovni jednotlivých článků, krátce po publikování. Nově navrhované metriky jako např. altmetrics obsahují řadu nových indikátorů, ale neposkytují jeden ucelený kvantifikovatelný indikátor. Výběr vhodných indikátorů a jejich váha je tak na uživateli a jeho konkrétních potřebách.

Naskýtá se otázka, zda alternativní metriky jsou vhodné pro publikace v Open Access repozitářích, včetně např. prací psaných v českém jazyce, kde lze očekávat nižší míru odezvy na sociálních sítích. Pro testování využil autor této práce statistik služby PlumX pro Pittsburskou univerzitu (16) a repozitář Pittsburské univerzity, konkrétně export metadat 1087 článků publikovaných v roce 2014 (17) ve formátu Dublin Core.

V doprovodné prezentaci na konferenci INFORUM je znázorněn součet hodnot jednotlivých metrik pro sledovaná data, tj. články v repozitáři Pittsburské univerzity publikované v roce 2014, rozdělené do kategorií Využití, Záložkování, Sociální média, Citace a Zmínky. Vybrané alternativní metriky byly analyzovány blíže, abychom získali lepší představu o distribuci a vhodnosti metrik pro hodnocení významu článků ve sledovaném repozitáři. Zpracované souhrnné statistiky zobrazuje Tabulka 2.

Tabulka 2 Statistické vyhodnocení metrik PlumX Pittsburské univerzity (zdroj: vlastní zpracování podle (16, 17))

	Pitts Downl.	Tweets	FB Shares	FB Likes	Bitly Clicks	FB Comments	Google+	Delicious
Zmínky	60 734	1 309	376	1 222	694	336	111	2
Nálezů	991	151	109	57	40	39	28	2
Zmínky / celkem	55,9	1,2	0,3	1,1	0,6	0,3	0,1	0,0
Nálezů / celkem	91,2%	13,9%	10,0%	5,2%	3,7%	3,6%	2,6%	0,2%
Zmínky / nálezů	61,3	8,7	3,4	21,4	17,4	8,6	4,0	1,0

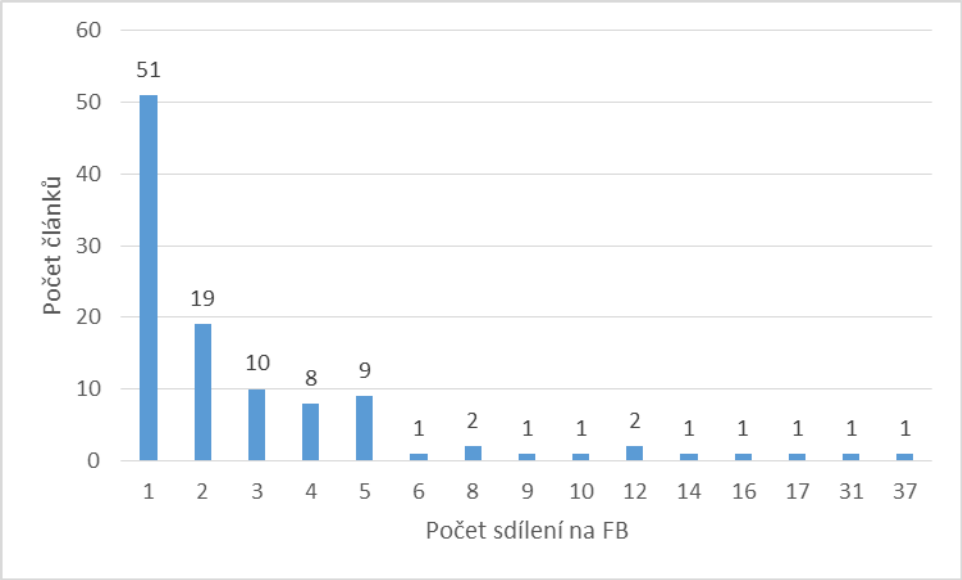
Řádek „Zmínky“ zobrazuje celkový počet všech nalezených zmínek o sledovaných článcích v konkrétní službě (tj. součet hodnot dané metriky za rok 2014). Jako „Nález“ je započítán pouze takový článek, o kterém je v PlumX evidována min. jedna zmínka v dané metrice.

Průměrný počet zmínek připadajících na jeden článek v repozitáři, údaj „Zmínky / celkem“, je vypočten jako podíl všech nalezených zmínek k počtu všech analyzovaných záznamů (1087).

Procentuální podíl článků, u kterých byla nalezena zmínka v konkrétní službě (tj. hodnota metriky je nenulová), uvádí řádek „Nálezů / celkem“. Průměrný počet zmínek u prací s pozitivním nálezem je uveden v řádku „Zmínky / nález“, tj. zde pomíjíme články, u kterých v databázi PlumX není evidováno využití v konkrétní službě a hodnota příslušné metriky je rovna nule.

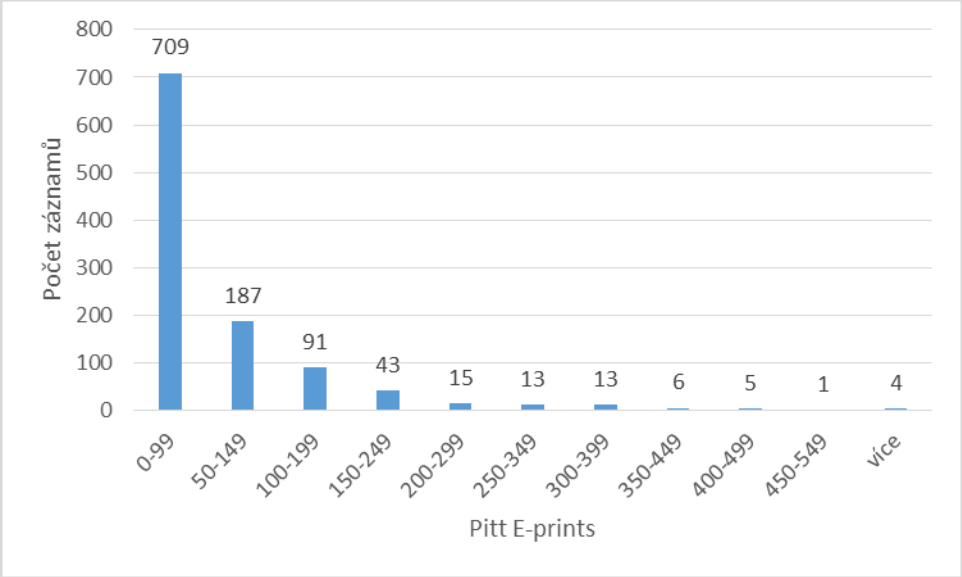
Pro analyzované články jsou nejméně využívané služby Delicious (nalezeny 2 články), sociální síť Google+ (111 zmínek o 28 článcích) a nástroj pro zkracování URL adres Bitly (694 kliknutí na zkrácenou URL 40 článků).

Ze sociálních sítí je nejvyužívanější Facebook, pro který je evidováno 379 sdílení u 109 prací. Jedno sdílení na Facebooku má v průměru před 3 označení Like a 1 komentář. Pravděpodobnost, že práce z repozitáře Pittsburské univerzity bude alespoň jednou sdílena na Facebooku je 10 %. Pokud však eliminujeme články sdílené na síti Facebook právě jednou, což může odpovídat nasdílení samotným autorem, získáme jen 58 nálezů namísto původních 109, což odpovídá 5 % všech analyzovaných článků. Histogram počtu sdílení na Facebooku znázorňuje Obrázek 6.



Obrázek 6 Histogram počtu sdílení článků na Facebooku (zdroj: autor)

V našem příkladu metrikou pokrývající největší počet článků je počet stažení plného textu z repozitáře univerzity. Histogram rozdělení počtu článků v repozitáři univerzity podle počtu stažení znázorňuje Obrázek 7.



Obrázek 7 Histogram metriky Pittsurgh E-Print

Metriky s nejvyššími hodnotami, rovnoměrněji distribuovanými mezi většinu testovaných článků, jsou počty zobrazení abstraktu (42 tisíc), PDF (67 tisíc) a HTML stránky s metadaty (343 tisíc). U těchto metrik je tedy předpoklad, že nebudou natolik ovlivněny samotným autorem, ale významem a atraktivností odpovídajících článků.

4 Závěr

V případě rozsáhlých Open Access repozitářů s významnými výsledky vědy a výzkumu nebo v případě publikace novinových článků na aktuální a nová témata se jako zajímavou náhradou klasických citačních metrik jeví tzv. alternativní metriky, založené na analýze ohlasů v online prostředí Internetu.

Podmínkou využití je atraktivnost dokumentu a jeho časté sdílení v sociálních sítích komunitou čtenářů.

U méně významných prací však nižší míra citovanosti prací na sociálních sítích může zapříčinit nízkou vypovídající schopnost konkrétní metriky. V našem testu by se mohlo jednat o málo využívané služby (např. Delicious nebo Google+) a o služby s nízkou hodnotou metriky (např. vice jak 5 sdílení na Facebooku má jen 12 z 1087 článků v testu, hrozí tak např. možnost umělého ovlivnění hodnoty metriky pro daný článek autorem).

Klasické metriky založené na impakt faktoru a citační analýze jsou vhodné pro hodnocení článků ve větším časovém horizontu po publikování. Metriky ukazující bezprostřední impakt článků v Open Access repozitářích, které by pomohly uživatelům rozpoznat významné či alespoň krátkodobě atraktivní články v repozitáři, by měly být založeny spíše na webometrikách, jako je např. počet stažení plného textu. V repozitářích je vhodné sbírat statistické podklady pro výpočet webometrických indikátorů, data deduplikovat a očistit o automatické přístupy robotů a vypočítané statistiky nabídnout uživatelům tak, abychom jim usnadnili jejich rozhodování o důležitosti a atraktivitě jednotlivých článků v repozitáři. Příkladem takovéto prezentace dat může být např. zobrazení nejstahovanějších prací za dané období nebo zobrazení historie počtu stažení plného textu na časové ose.

Míru užití záznamů je nyní možné vyhodnocovat pouze na úrovni jednoho repozitáře, ani to však není u repozitářů v ČR plně využíváno. Protože hrozí odlišná interpretace webových logů jednotlivých repozitářích podle použité metodiky výpočtu (např. odlišný způsob počítání návštěv a filtrování přístupu robotů), nejde výsledky mezi jednotlivými repozitáři srovnávat.

Pro zpracování a vyhodnocení statistik nad Open Access repozitáři v ČR autor navrhuje realizaci společného projektu na vybudování centrálního agregačního bodu, který by sbíral a jednotně vyhodnocoval statistiky pomocí v textu popsaných metod a standardů. Vzorová implementace vhodná i pro podmínky repozitářů v ČR byla popsána na projektu IRUS-UK.

Použitá literatura

1. LOZANO, George A. a Vincent LARIVIÈRE, Yves GINGRAS. The weakening relationship between the Impact Factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*. 8. 10. 2012. DOI: 10.1002/asi.22731. Dostupné také z: http://arxiv.org/abs/1205.4328
2. AMIN, Mayur a Michael MABE. ImpactFactors: Use and Abuse. *Perspectives in Publishing*. 2000, č. 1, s. 1-6. Reissued with minor revisions 11. 2007. Dostupné z: http://cdn.elsevier.com/assets/pdf_file/0014/111425/Perspectives1.pdfPLOS ONE
3. PLOS ONE [online]. PLOS, 2015 [cit. 4. 5. 2015]. ISSN eISSN-1932-620. Dostupné z: http://www.plosone.org/
4. The OpenAIRE Consortium. Usage Statistics from repositories. *OpenAire: open access infrastructure for reserach in Europe* [online]. 2012 [cit. 5. 2. 2015]. Dostupné z: https://www.openaire.eu/content
5. PRIEM, Jason aj. Altmetrics: a manifesto. In: *Altmetrics* [online]. Verze 1.01, 28. 9. 2011 [cit. 4. 5. 2015]. Dostupné z: http://altmetrics.org/manifesto/
6. IMPACTSTORY. Carl Boettiger: Is your phylogeny informative? Measuring the power of comparative methods. *Impactstory* [online]. 2015 [cit. 18. 2. 2015]. Dostupné z: https://impactstory.org/CarlBoettiger/product/t2q1a39jt3kythditpt30uhu/metrics

7. Homepage. PLUM ANALYTICS. *PlumX* [online]. 2015 [cit. 4. 5. 2015]. Dostupné z: https://plu.mx/
8. Counter Online Metrics: The COUNTER Code of Practice for e-Resources: Release 4. *COUNTER* [online]. 4. 2012 [cit. 4. 5. 2015]. Dostupné z: http://www.projectcounter.org/r4/COPR4.pdf
9. COUNTER. COUNTER Code of Practice for Articles. In: *COUNTER* [online]. 4. 2012 [cit. 4. 5. 2015]. Dostupné z: http://www.projectcounter.org/documents/counterart_cop_MAR2014.pdf
10. SHEPHERD, Peter a Paul NEEDHAM. *Publisher and Institutional Repository usage Statistics: The PIRUS2 Project: final report* [online]. Cranfield: Cranfield University, 6. 10. 2011 [cit. 4. 5. 2015]. Dostupné z: http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-download_wiki_attachment.php?attId=170&download=y
11. VERHAAR, Peter. KE Usage Statistics Guidelines. *SUFR wiki* [online]. Verze 1.0, 18. 5. 2010 [cit. 4. 5. 2015]. Dostupné z: http://wiki.surfnet.nl/display/standards/KE+Usage+Statistics+Guidelines
12. IRUS-UK. *IRUS-UK* [online]. 2014 [cit. 1. 5. 2015]. Dostupné z: http://www.irus.mimas.ac.uk/
13. NEEDHAM, Paul. IRUS UK: Making ETDs count in UK repositories. In: *ETD 2014*. University of Lecester, 2014. Dostupné také z: https://www.youtube.com/watch?v=flocun3wAZU
14. The PIRUS Code of Practice for recording and reporting usage at the individual article level. In: *PIRUS* [online]. Verze 1. říjen 2013 [cit. 4. 5. 2015]. A COUNTER Standard. Dostupné z: http://www.projectcounter.org/documents/Pirus_cop_OCT2013.pdf
15. The Tracker protocol V3.1. *IRUS-UK* [online]. 22. 4. 2014 [cit. 25. 4. 2015]. Dostupné z: http://www.irus.mimas.ac.uk/help/toolbox/TrackerProtocol-V3-2014-04-22.pdf
16. PlumX / Pitt. In: *PlumX* [online]. 2015 [cit. 2. 5. 2015]. Dostupné z: https://plu.mx/pitt/g/?artifact_tab=ARTICLE&custom_filter=&custom_fil_g=&r_year=2014
17. D-Scholarship @ Pitt. *University of Pittsburgh* [online] 2014 [cit. 2. 5. 2015] Dostupné z: http://d-scholarship.pitt.edu/view/year/2014.html